

Deep origin of eukaryotes outside Heimdallarchaeia within Asgardarchaeota

<https://doi.org/10.1038/s41586-025-08955-7>

Jiawei Zhang¹, Xiaoyuan Feng², Meng Li², Yang Liu², Min Liu³, Li-Jun Hou^{1✉} & Hong-Po Dong^{1✉}

Received: 18 September 2024

Accepted: 28 March 2025

Published online: 07 May 2025

Open access

 Check for updates

Research on the morphology, physiology and genomics of Asgard archaea has provided valuable insights into the evolutionary history of eukaryotes^{1–3}. A previous study suggested that eukaryotes are nested within Heimdallarchaeia⁴, but their exact phylogenetic placement within Asgard archaea remains controversial^{4,5}. This debate complicates understanding of the metabolic features and timescales of early eukaryotic ancestors. Here we generated 223 metagenome-assembled nearly complete genomes of Asgard archaea that have not previously been documented. We identify 16 new lineages at the genus level or higher, which substantially expands the known phylogenetic diversity of Asgard archaea. Through sophisticated phylogenomic analysis of this expanded genomic dataset involving several marker sets we infer that eukaryotes evolved before the diversification of all sampled Heimdallarchaeia, rather than branching with Hodarchaeales within the Heimdallarchaeia. This difference in the placement of eukaryotes is probably caused by the previously underappreciated chimeric nature of Njordarchaeales genomes, which we find are composed of sequences of both Asgard and TACK archaea (Asgard's sister phylum). Using ancestral reconstruction and molecular dating, we infer that the last Asgard archaea and eukaryote common ancestor emerged before the Great Oxidation Event and was probably an anaerobic H₂-dependent acetogen. Our findings support the hydrogen hypothesis of eukaryogenesis, which posits that eukaryotes arose from the fusion of a H₂-consuming archaeal host and a H₂-producing protomitochondrion.

The origin of eukaryotic cells has long remained a mystery⁶. Asgard archaea contain more genes encoding eukaryotic signature proteins (ESPs) than other archaea and are deemed to be the closest relatives of eukaryotes^{4,5,7–9}. Isolation and cultivation of the first Asgard archaeon (*'Candidatus' Prometheoarchaeum (Ca. P.) syntrophicum' MK-D1*), belonging to Lokiarchaeia, revealed that the archaeon has long and branching protrusions and is able to degrade amino acids syntrophically¹. These morphological and metabolic features provide new insights into the episodes of eukaryogenesis. For example, it is assumed that cell protrusion may mediate cell–cell interactions, leading to engulfment of its partner and gradual formation of mitochondrion. Subsequently, the second Asgard archaeon (*'Candidatus' Lokiarchaeum (Ca. L.) ossiferum'*) was recently enriched, and cryo-electron tomography and immunostaining revealed the presence of the actin-like cytoskeletal filament and Lokiactin expression². It is inferred that the Lokiactin may have served as a scaffold for the Asgard archaeal cell and played a role in maintenance of cell shape, cell division and molecular trafficking, similar to actin filaments in eukaryotic cells^{10,11}. The results and hypotheses implicate that the eukaryotic nuclear lineages evolved from within the archaea.

Although some eukaryotic-like features have been identified in pure culture or enrichment of Asgard archaea, with such a limited number of strains, it is challenging to infer the exact phylogenetic position, timescales and metabolic characteristics of early eukaryotic ancestors. In contrast, phylogenomic analyses based on a wealth of archaeal genomic data can compensate for these shortcomings¹². By using super-tree and coalescent methods using more than 3,000 gene families in archaea and eukaryotes, the Heimdallarchaeota have been proposed as the closest relatives of eukaryotes¹³. However, the study used only six Asgard archaeal genomes and failed to represent the currently expanded diversity of Asgard archaea. Recently, two studies obtained inconsistent results revolving around the relationship between eukaryotes and Asgard archaea based on an expanded genomic sampling of Asgard archaea. One suggested that eukaryotes are more likely to have branched from within Asgard archaea as a sister group to the Heimdallarchaeota–Wukongarchaeota branch or are a deeper branch within archaea⁵, whereas another proposed that eukaryotes branch as a sister lineage to Hodarchaeales within Heimdallarchaeia⁴. The former used only 29 concatenated markers to infer phylogenies, whereas the latter used non-ribosomal proteins as markers in combination with

¹State Key Laboratory of Estuarine and Coastal Research, Yangtze Delta Estuarine Wetland Ecosystem Observation and Research Station, Ministry of Education, East China Normal University, Shanghai, China. ²Archaeal Biology Center, Synthetic Biology Research Center, Shenzhen Key Laboratory of Marine Microbiome Engineering, Key Laboratory of Marine Microbiome Engineering of Guangdong Higher Education Institutes, Institute for Advanced Study, Shenzhen University, Shenzhen, China. ³Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai, China. ✉e-mail: ljhou@sklec.ecnu.edu.cn; hpdong@sklec.ecnu.edu.cn

extensive exclusion of the sites from the alignments. The exact relationship between eukaryotes and archaea remains to be further clarified.

Here we present 223 new Asgard archaeal genomes, including the identification of 16 additional order-, family- or genus-level lineages, recovered from metagenomic samples generated from 14 sites across coastal wetlands of China, which significantly expand the diversity of Asgard archaea. By analysing the expanded genomic sampling of Asgard archaea leveraging sophisticated phylogenomic approaches, including recoding of alignments, use of complex site-heterogeneous evolution models in maximum likelihood (ML) and Bayesian inferences and reduction of rate heterogeneity, we robustly place eukaryotes within the Asgard archaea as a sister clade to Heimdallarchaeia. By applying ancestral reconstruction together with molecular dating, the timescale and metabolic traits of the last Asgard archaea and eukaryote common ancestor (LAECA) were delineated and found to differ significantly from previous studies^{1,4}.

Expanded diversity of Asgard archaea

The increase in genomic diversity of Asgard contributes to resolving the evolutionary relationship between eukaryotes and Asgard archaea^{4,5,7}. To this end, we collected 40 sediment samples from salt marsh and mangrove wetlands across China and performed metagenomic sequencing (Supplementary Table 1 and Supplementary Fig. 1a). After *de novo* assembly and binning of scaffolds, we reconstructed 11,878 metagenome-assembled genomes (MAGs) (Supplementary Fig. 1b). Of these MAGs, 223 belonged to the Asgard archaea, with greater than 70% completeness and less than 10% contamination (Supplementary Table 2 and Supplementary Fig. 1c). These MAGs were combined with 395 publicly available genomes with high quality. By using dRep¹⁴ to select representatives at species level (ANI at least 95%), we finally obtained a set of 411 Asgard genomes, of which 136 are derived from this study (Supplementary Table 2). Among these, two are from cultured Asgard archaea (*Ca. P. syntrophicum* and *Ca. L. ossiferum*) while the remaining genomes are MAGs recovered from environmental samples (Fig. 1). These genomes have a mean completeness of 85.3% and a mean contamination of 3.6%.

To resolve the phylogenetic position of these clades, we inferred an ML tree of a concatenated set of 53 archaeal-specific marker proteins in the Genome Taxonomy Database (GTDB)¹⁵ (GTDB.ar53) for a phylogenetically diverse set of archaeal genomes (411 Asgard archaea, 51 DPANN (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota) archaea, 47 Euryarchaea and 70 TACK (Thaumarchaeota, Aigarchaeota, Crenarchaeota, Korarchaeota) archaea representatives, Supplementary Table 3). The tree confirmed the existence of ten class-level Asgard archaeal lineages (Fig. 1). Based on the relative evolutionary distance¹⁶, we identified two putative order-level lineages in Odinarchaeia and Heimdallarchaeia (relative evolutionary distance less than 0.53), which were provisionally named Yangjianarchaeales and Wenzhongarchaeales. In addition, four new family-level clades (represented by six MAGs) and ten new genus-level clades (represented by 18 MAGs) were identified within Sifarchaeia, Hodarchaeales, Hermodarchaeia and Lokiarchaeia. In six out of the ten Asgard archaea classes, we presented 108 species-level new genomes (Fig. 1). Currently, the placement of the recently proposed Njordarchaeales in archaeal phylogeny is intensely debated^{4,17,18}. Recently, it was placed within Heimdallarchaeia in the tree inferred using new alternative markers (NM57), forming a clade with Gerdarchaeales, Kariarchaeaceae and Heimdallarchaeaceae⁴. We selected 12 sets of frequently used marker proteins from previous studies (Supplementary Table 4)^{4,5,7,8,13,19–23} that are conserved in most of archaeal clades (Methods). ML phylogenomic analyses of these marker sets, based on a set of 579 archaeal representatives including the enlarged genomic sampling of Asgard archaea, were performed under a sophisticated evolutionary model (LG + C60 + F + G + PMSF). These results showed

that, except for in the NM57 tree, Njordarchaeales was placed within the TACK superphylum and branched as a sister lineage to Korarchaeota with high support in the other 11 trees (Supplementary Fig. 2). Given that Njordarchaeales often branched with eukaryotes in previous phylogenomic analyses^{4,18}, it is essential to identify the factors underlying the conflicting topologies among different trees for position of Njordarchaeales within archaea.

Revisit of Njordarchaeales placement

Eme et al.⁴ argued that the monophyly of Njordarchaeales and Korarchaeota in the RP56 tree may be caused by high compositional similarity in RP56 ribosomal marker protein sequences, which is connected with thermophilic lifestyle of the two lineages. Based on this hypothesis, they proposed a new marker dataset consisting of 57 proteins of archaeal origin that excluded ribosomal proteins to infer the phylogeny of archaea. The NM57 markers are selected from a set of 200 archaeal conserved marker proteins (NM200) identified by Petitjean et al.²¹ and most of them are functional proteins involved in metabolic and cellular processes. We observed that, despite the absence of ribosomal proteins, the ML tree of the NM200 dataset positioned Njordarchaeales as a sister clade to Korarchaeota within the TACK superphylum (Supplementary Fig. 2k). This suggests that the clustering of Njordarchaeales and Korarchaeota is not necessarily related to ribosomal proteins. The inconsistent topologies may be attributed to the differences in the number and nature of amino acid positions in the alignments of the two datasets. In contrast to the NM57 topology, the NM200 tree, with more amino acid sites, is more likely to represent the true evolutionary relationship between Njordarchaeales and other archaea.

In phylogenomic analysis, the even distribution of marker proteins in the different archaeal phyla or clades is crucial for clarifying the phylogenetic relationships between archaeal clades. To this end, we integrated the 12 sets of frequently used marker proteins from previous studies (Supplementary Table 4) and selected 67 markers (S67) that are conserved across all sampled archaeal genomes (identified in at least 60% of representatives of each of the archaeal clades; Methods). The S67 marker set was comprised of 39 ribosomal proteins and 28 functional proteins involved in diverse cellular activities (Supplementary Table 4). ML phylogenomic analyses of the S67 dataset under sophisticated evolutionary models confirmed Njordarchaeales as a sister to Korarchaeota and resolved all nodes of the Asgard archaea with high support (Bootstrap values at least 95) (Extended Data Fig. 1).

Given that concatenation of large amount of data may strengthen non-phylogenetic signal and lead to artifactual trees with high support, we next evaluated the effect of data exclusion on phylogenetic inference by removing mutationally saturated or rate-biased sites^{24–27}. We tracked the evolution of bootstrap support for the monophyly of either (1) Njordarchaeales and Korarchaeota or (2) Njordarchaeales and Heimdallarchaeia in trees inferred from the S67 alignments using complex site-heterogeneous evolution model, as fast-evolving sites are progressively removed. The grouping of Njordarchaeales and Korarchaeota consistently received strong support until 80% or more of the fastest-evolving sites were removed, whereas the monophyly of Njordarchaeales and Heimdallarchaeia was never supported (Extended Data Fig. 2a). The significant loss of phylogenetic signal probably accounted for the decline in support observed after the removal of 80% or more of the fastest-evolving sites.

Nevertheless, it is observed that the incremental exclusion of the fastest-evolving sites from the NM57 alignments also strongly supported the grouping of Njordarchaeales and Heimdallarchaeia (Extended Data Fig. 2b). To identify the cause of this effect, we determined the taxonomic profiles of contigs of the ten Njordarchaeales representatives using CAT^{28,29} and MMseqs2 (refs. 30,31) taxonomy tools. Although the two tools use different algorithms, they generated comparable results. In each genome, approximately 24–60% of the contigs

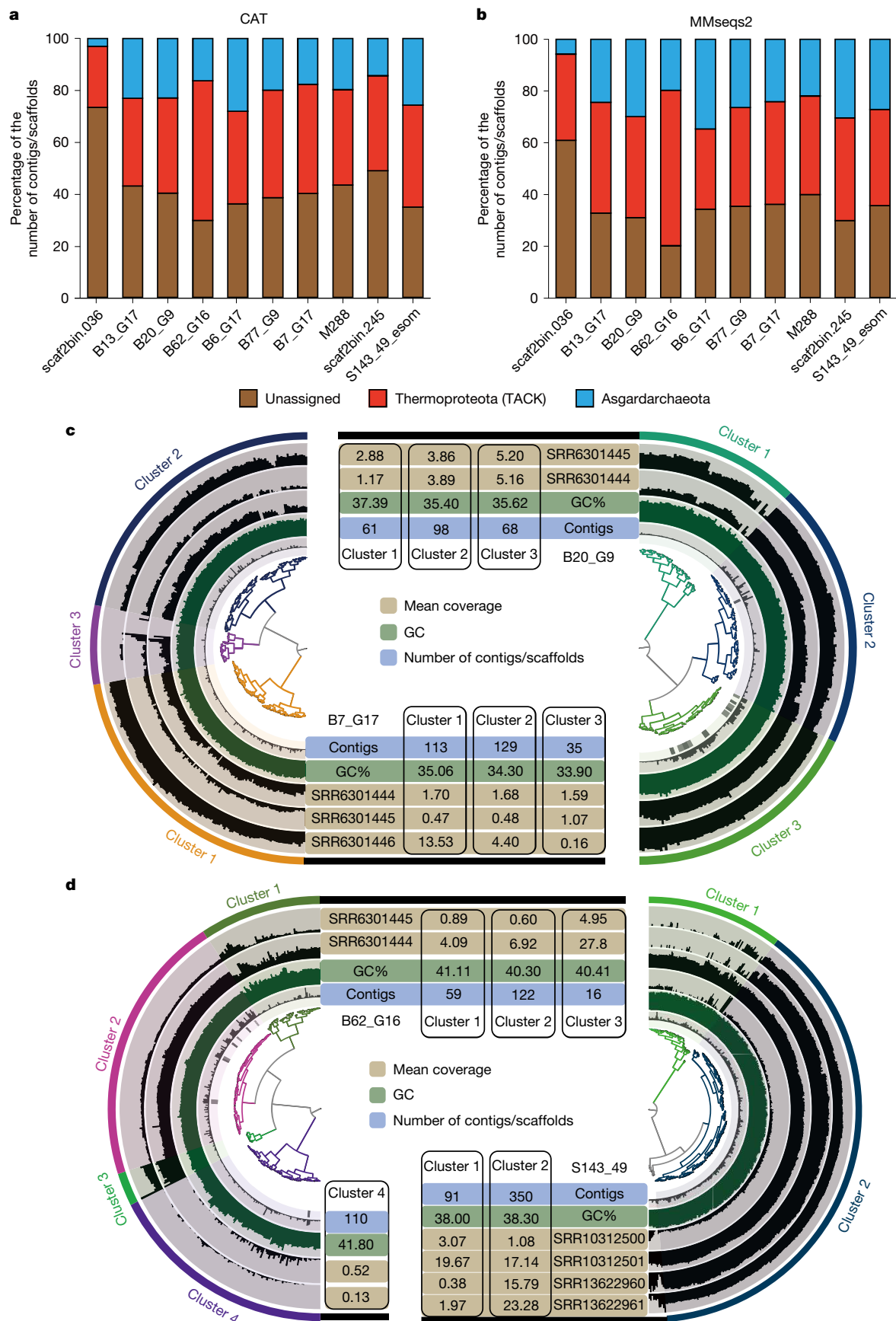


Fig. 2 | Taxonomic profiles and clustering of contigs/scaffolds in Njordarchaeales genomes. a,b, Percentage of contigs/scaffolds assigned to Thermoproteota (formerly known as the TACK superphylum) or Asgardarchaeota based on count for each of the ten Njordarchaeales representatives; classification of contigs/scaffolds was determined using CAT (**a**) or MMseqs2 (**b**) tools. **c,d,** Hierarchical clustering of contigs/scaffolds in Njordarchaeales B7_G17 and B20_G9 (**c**) and B62_G16 and S143_49 (**d**) genomes based on their sequence composition and differential mean coverage across different metagenomes.

MAGs may represent chimeric assemblies derived from two or three distinct populations. Compared with other datasets, The NM57 dataset may harbour more robust phylogenetic signals that support the Asgard part of these MAGs, possibly because the NM57 markers were selected based on their distribution across Asgard archaea⁴. The taxonomic profiles of Njordarchaeales and its phylogenetic trees inferred using ribosomal proteins-containing marker sets (Fig. 2, Supplementary Figs. 2 and 4 and Supplementary Discussion) revealed that the bulk of these MAGs appeared to be affiliated with the TACK archaea. The inclusion of such chimeras in tree reconstruction can be expected to impact the evolutionary position of eukaryotes within archaea.

Eukarya emerged outside Heimdallarchaeia

Our expanded Asgard archaeal genomes help to resolve the placement of eukaryotes among archaea. To this end, we added 14 commonly used eukaryotic taxa^{4,5,7} (Supplementary Table 3) into the S67 dataset to generate a supermatrix: ES67. The ML tree of the ES67 dataset placed eukaryotes within the TACK superphylum as a sister clade to Njordarchaeales and Korarchaeota (Supplementary Fig. 5a), which had never been observed previously in previous studies^{4,17}. This may be caused by the high compositional similarity between the Asgard archaeal sequences in the Njordarchaeales genomes and eukaryotic sequences. To mitigate the effect of distant outgroups on the placement of eukaryotes, we removed DPANN, Euryarchaea, Korarchaeota and Njordarchaeales sequences from the ES67 dataset while retaining 50 representatives of TACK archaea as the outgroup; alternatively, all outgroup sequences were omitted. The resulting supermatrices were named tES67 and AsES67, respectively. ML analysis of the tES67 dataset showed that eukaryotes branched with Heimdallarchaeia with high support (Heimdallarchaeia-sister) (Fig. 3a and Supplementary Fig. 5b). The unrooted ML phylogeny of Asgard archaea and eukaryotes on the basis of the AsES67 dataset also strongly supported the Heimdallarchaeia-sister topology (Supplementary Fig. 5c).

Based on the tES67 and AsES67 datasets, we examined the effect of rate heterogeneity on the position of eukaryotes by progressively removing fast-evolving sites of the two datasets. Along the decreasing heterogeneity gradient, ML phylogenies were computed to assess the support for Heimdallarchaeia-sister and the monophyly of eukaryotes and Hodarchaeales (Hodarchaeales-sister). Support for Hodarchaeales-sister was never observed, whereas support for Heimdallarchaeia-sister fluctuated as rate heterogeneity decreased (Extended Data Fig. 4). We inferred that the fluctuation in support for Heimdallarchaeia-sister may be ascribed to the dynamic relationship between the loss of phylogenetic signals and the level of rate heterogeneity. Although Bayesian inferences of the SR4-recoded³⁴ tES67 and AsES67 datasets displayed the Heimdallarchaeia-sister topology, support was weak (posterior probability (PP) support values of 0.69 and 0.5, respectively) (Fig. 3a and Supplementary Fig. 6). The results suggest that the S67 marker set does not contain enough phylogenetic signal to resolve the node between eukaryotes and Asgard clades.

To address the issue, we relaxed our marker selection criterion. Based on conservation of marker proteins in the expanded Asgard genomes, we selected two additional marker sets from previously reported marker proteins that are conserved in archaea^{4,5,7,13,19,21}. The first marker set consisted of 97 markers (S97) that were found in at least 60% of representatives of each of the Asgard clades, whereas the second marker set included 150 markers (S150) that were identified in at least 80% of Asgard archaeal genomes (Supplementary Table 4). ML trees inferred from the untreated S97 dataset and its fastest-evolving site-excluded subsets based on Asgard archaea and eukaryotic taxa provided significant support for the Heimdallarchaeia-sister topology (bootstrap support greater than 95) (Fig. 3b,c). In contrast, an ML tree inferred from the untreated S150 dataset strongly supported the clustering of eukaryotes and Hodarchaeales (Extended Data Fig. 5a).

Given that the larger amount of missing data contained in the S150 supermatrix could bias the estimation of the site evolutionary rates, a site-by-site desaturation strategy was also applied to the S150 supermatrix. We observed that support for the Hodarchaeales-sister was replaced by support for the Heimdallarchaeia-sister after the 40% fastest-evolving sites were removed (Extended Data Fig. 5b). The grouping of eukaryotes with Heimdallarchaeia was also confirmed by the Bayesian inferences of the SR4-recoded S97 and S150 datasets (PP values of 1.0 and 0.8, respectively; Fig. 3b, Supplementary Fig. 7).

In the study by Eme et al.⁴, based on the NM57 marker set, eukaryotes were positioned within Heimdallarchaeia, as a sister lineage to Hodarchaeales. We redid phylogenomic analyses of the NM57 dataset using the expanded Asgard archaea and eukaryotes. Although the ML tree of the untreated NM57 supported the Hodarchaeales-sister topology (Supplementary Fig. 8a), the support for Hodarchaeales-sister decreased dramatically, whereas support for Heimdallarchaeia-sister rose markedly as fastest-evolving sites were progressively reduced (Fig. 3c). The Bayesian inferences of the SR4-recoded³⁴ NM57 dataset also confidently placed eukaryotes outside Heimdallarchaeia (PP, 1.0) (Supplementary Fig. 8b). The result indicated that the relationship between eukaryotes and Asgard archaea reported⁴ was probably artifactual. Additional tests using vertically evolving markers that support the Heimdallarchaeia-sister topology are delineated in Supplementary Information (Supplementary Table 4 and Supplementary Figs. 9–12). In addition, we examined the distribution of ESPs in the expanded sampling of Asgard archaea and found that the previously delineated ESPs specific to Hodarchaeales were also present in other Asgard clades (Supplementary Results and Discussion and Supplementary Figs. 13–16).

Taken together, our phylogenetic analyses indicate that eukaryotes may have evolved before the diversification of all sampled Heimdallarchaeia. The results were consistent with one of scenarios inferred using the 29-marker tree reported previously by Liu et al.⁵, but challenged the recently proposed Hodarchaeales-sister hypothesis⁴. Compared with the study of Liu et al.⁵, our results were more robust as they were generated using an expanded genomic sampling of Asgard archaea, in combination with exclusion of outgroups, removal of fast-evolving sites and application of complex site-heterogeneous evolution models and several marker sets. Our finding helps to infer the nature of the eukaryotic ancestors.

Timescale and traits of the LAECA

We used MCMCTree³⁵ to estimate the divergence times of key nodes within the Asgard phylum using a relaxed molecular clock model. An archaeal root age range of 4.29–3.8 billion years ago (Ga) and uniform age priors for the other fossil calibrations were applied in the molecular dating analysis (Supplementary Table 5). The last Asgard archaea common ancestor (LASCA) emerged at around 3.38 Ga (confidence interval (CI), 3.72–3.06 Ga) in the Archean (Supplementary Fig. 17). The estimate for the age of the Heimdallarchaeia and Lokiarchaeia common ancestor was about 3.16 Ga (CI, 3.54–2.82 Ga) in the Mesoproterozoic. As the sister of eukaryotes, crown Heimdallarchaeia diverged 3.12–2.26 Ga, predating the Great Oxidation Event³⁶. The result indicates that eukaryotic stem lineages evolved before the Great Oxidation Event, which is basically consistent with the timescale of stem eukaryotes reported by Betts et al.³. Recently, based on the evolutionary history of ATP synthases, it was suggested that eukaryotes diverged from Hodarchaeales within Heimdallarchaeia 2.67–2.19 Ga³⁷. This partially overlaps in time with our estimates for divergence of Heimdallarchaeia and eukaryotes.

We used a probabilistic gene–tree species-tree reconciliation method³⁸ (amalgamated likelihood estimation (ALE)) to infer ancestral events and gene content on key nodes of the Asgard archaeal species tree. To refine the metabolic reconstruction, a subset of the aforementioned archaeal genomes with more than 80% completeness and less than 5% contamination was selected for ALE analyses. This subset, which

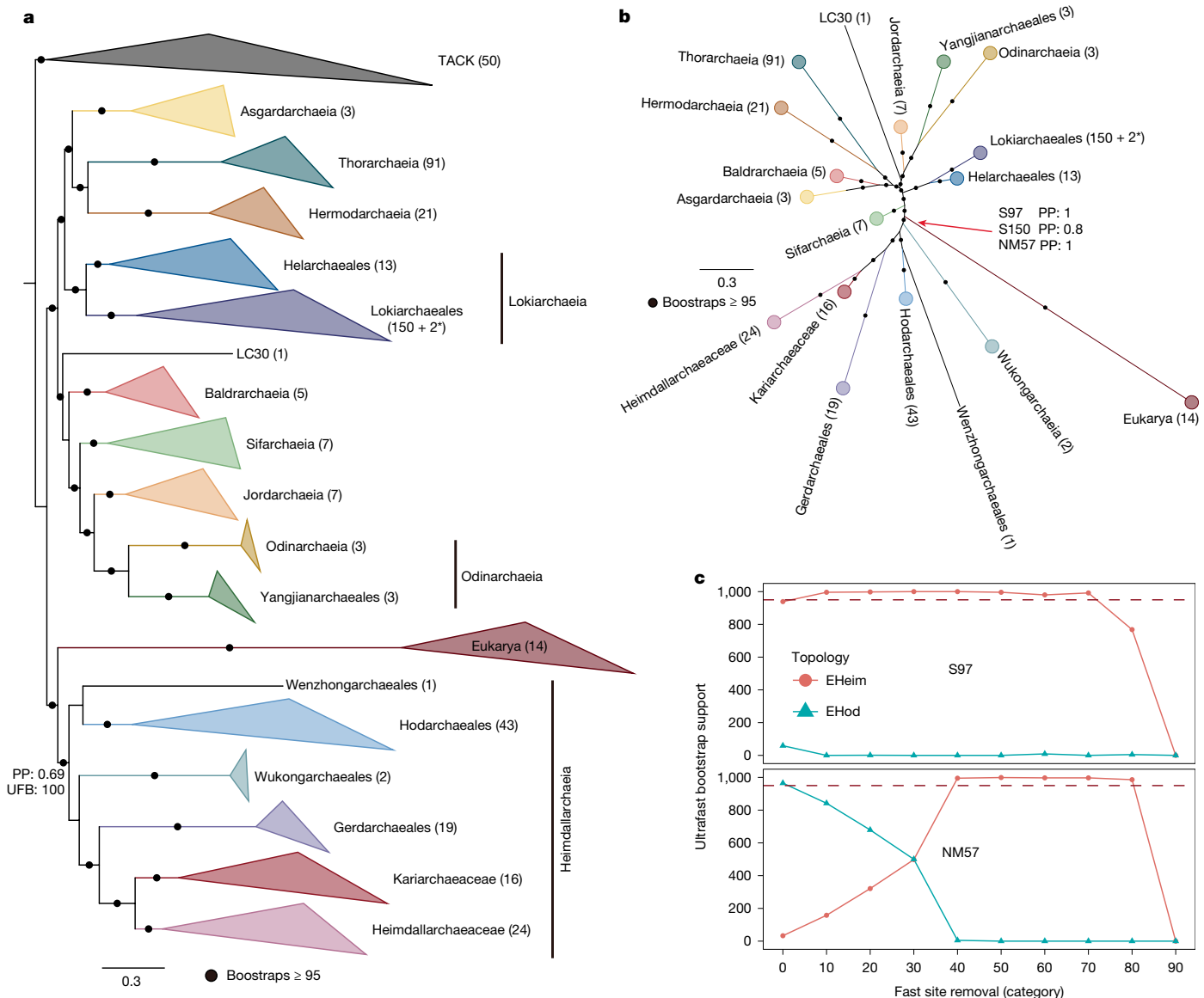


Fig. 3 | Phylogenomic analyses of several sets of concatenated marker proteins, showing the placement of eukaryotes relative to genomically sampled Asgard archaea. a, ML phylogenetic analysis of 67 concatenated marker proteins based on 461 archaeal taxa and 14 eukaryotes (inferred using IQ-TREE under LG + C60 + F + G + PMSF model), using 50 TACK archaea as the outgroup (tES67 alignment; 13,348 sites). The number below Lokiarchaeales represents 150 MAGs and the genomes of two cultured Asgard strains. Meanwhile, the alignment was SR4-recoded and its Bayesian inference was performed using the CAT + GTR model (Supplementary Fig. 6a; two chains; 50,000 generations). PP support for the node of eukaryotes and the closest Asgard relatives is shown (0.69). **b**, ML phylogenetic analysis of 97 concatenated

marker proteins based on 411 Asgard archaeal taxa and 14 eukaryotes (inferred using IQ-TREE under the LG + C60 + F + G + PMSF model) (S97 alignment; 20,067 sites). This tree is unrooted. In addition, Bayesian inferences of the SR4-recoded S97, S150 and NM57 supermatrices were performed using the CAT + GTR model (Supplementary Figs. 7 and 8b). PP support for the node of eukaryotes and the closest Asgard relatives is shown (PP = 1 for S97, PP = 0.8 for S150 and PP = 1 for NM57). **c**, Evolution of bootstrap support for the grouping of eukaryotes with Heimdallarchaeia (Heimdallarchaeia-sister, EHeim) or Hodarchaeales (Hodarchaeales-sister, EHod) in phylogenetic trees inferred from S97 and NM57 datasets, as the fastest-evolving sites were progressively removed.

included 235 Asgard archaea, 47 Euryarchaea and 60 TACK archaea, had a mean completeness of 88.5% and a mean contamination of 2.7% (Supplementary Tables 2 and 3). In the species tree, a gradual increase in ancestral gene content was observed during the evolution from LasCA to Heimdallarchaeia ancestor (Supplementary Fig. 18). Along the evolutionary route, rates of gene loss also increased gradually, implying that a large amount of gene gain occurred before Heimdallarchaeia radiation.

Next, we reconstructed metabolic traits of key ancestors of Asgard archaea based on presence probability of genes in nodes reported by ALE (Supplementary Table 6) and gene frequencies across Asgard

archaeal genomes (Supplementary Table 7). Recent studies on two cultured Asgard archaea, *Ca. P. syntrophicum* and *Ca. L. ossiferum*, have shown their ability to degrade amino acids anaerobically by syntrophy^{1,2}. Based on these physiological features, it was inferred that the LasCA was probably an amino-acid-degrading anaerobe that produced H₂ and fatty acids as byproducts³⁹. However, our analysis revealed that the two cultivated Asgard strains and their three closely related MAGs lacked most of the key genes for the H₄MPT methyl branch of the archaeal Wood–Ljungdahl pathway⁴⁰ (WLP), but encoded three key enzymes in the H₄F methyl branch of bacterial WLP: formyltetrahydrofolate synthetase (*fhs*), methylenetetrahydrofolate dehydrogenase

(*fold*) and methylenetetrahydrofolate reductase (*metF*) (Supplementary Fig. 19). This partial H₄F branch can convert glycine, serine and histidine to formate that mediates interspecies electron transfer¹ (Supplementary Fig. 20). In contrast, the LAsCA was inferred to harbour a complete archaeal WLP but lacked Fhs in the bacterial WLP (Fig. 4a, Supplementary Tables 6 and 7 and Supplementary Figs. 19 and 20). It also encoded various NiFe hydrogenases (Supplementary Table 6 and Supplementary Fig. 21). These results suggest that the LAsCA was a H₂-dependent chemolithoautotroph. Abundant CO₂ in the Archean ocean⁴¹ and marine H₂ from serpentinization⁴² would have provided ample substrates for these organisms.

Lokiarchaea included both Helarchaeales and Lokiarchaeales. All members of Helarchaeales (13 MAGs) are found to encode a methyl-CoM reductase-like enzyme (Supplementary Table 7) that is similar to that found in butane-oxidizing archaea^{43,44}. ALE analysis inferred the presence of the methyl-CoM reductase-like enzyme in Lokiarchaea ancestor (Supplementary Table 6), suggesting that this ancestor had the potential to degrade hydrocarbon anaerobically by coupling the β -oxidation of fatty acids with reverse WLP (Fig. 4a and Supplementary Fig. 20).

Contrary to the result by Eme et al.⁴, our inference revealed that the Heimdallarchaea and Hodarchaeales ancestors also possessed a complete archaeal WLP pathway (Fig. 4a, Supplementary Fig. 20 and Supplementary Table 6). We found that, in the MAGs of Heimdallarchaea (including three Hodarchaeales genomes) used by Eme et al.⁴, genes encoding several key enzymes in the archaeal WLP are missing. However, we presented 22 new genomes of Hodarchaeales, representing two new families and seven new genera, in which several genomes or lineages contained a complete archaeal WLP (Fig. 4b). As a basal-branching group in Heimdallarchaea, the presence of archaeal WLP in Hodarchaeales supported an autotrophic lifestyle of crown Heimdallarchaea. Except for the WLP, the crown Heimdallarchaea was inferred to encode a phosphotransacetylase (*pta*) and an acetate kinase (*ack*) that are required for the conversion from acetyl-CoA to acetate, as well as an energy-converting hydrogenase (*ech*) that can potentially serve as an ion-translocating chemiosmotic coupling site, and an ATP synthase (Supplementary Tables 6 and 7 and Supplementary Fig. 20). Gene frequency analysis revealed that *ack* genes were found exclusively in members of Heimdallarchaea (Supplementary Fig. 21a). These data inspired the inference that Heimdallarchaea ancestor was an anaerobic H₂-dependent acetogen, as shown in a previous study⁴⁵.

As a parent of crown Heimdallarchaea, the LAECA may also be a H₂-dependent acetogenic archaeon. Our metabolic evidence supported the hydrogen hypothesis, which posits that eukaryogenesis was mediated initially by a syntrophic partnership between a H₂-consuming archaeal host and a H₂-producing protomitochondrion living under anoxic conditions, with the protomitochondrion being a facultative aerobe^{46,47}. We identified an incomplete cytochrome *bd-I* ubiquinol oxidase and an incomplete cytochrome *o* ubiquinol oxidase in 38 members of Hodarchaeales and 16 members of Kariarchaeaceae (Supplementary Fig. 21b and Supplementary Table 7). However, only two subunits of the cytochrome *o* ubiquinol oxidase were inferred to be present in the Heimdallarchaea ancestor (Supplementary Fig. 20 and Supplementary Table 6), which may have been acquired through HGT from a bacterial partner. These incomplete complexes probably lacked the capacity to transport electrons or reduce molecular oxygen to water. The inferred presence of extremely O₂ sensitive enzymes, such as acetyl-CoA decarboxylase/synthase (*cdh*), 2-oxoacid: ferredoxin oxidoreductases (*kor*, *vor* and *por*) and NiFe hydrogenases in the Heimdallarchaea ancestor (Supplementary Fig. 20 and Supplementary Table 6 and 7), together with the anoxic deep ocean environment at that time⁴⁶, suggests that this ancestor must have grown anaerobically.

In summary, we used sophisticated phylogenomic methods in combination with a significantly expanded sampling of Asgard archaea to re-evaluate relationships among Asgard archaea, Njordarchaeales and eukaryotes. We revealed that the Njordarchaeales genomes probably

represent chimeric assemblies comprising mainly TACK and Asgard archaeal sequences. These weird assembled genomes would significantly affect phylogenetic position of eukaryotes in archaea. After excluding outgroups, using several marker sets, eukaryotes were placed confidently within Asgard archaea as a sister to Heimdallarchaea instead of being nested within Heimdallarchaea branching with Hodarchaeales. Ancestral reconstructions inferred that the host lineage at eukaryotic origin was an anaerobic, H₂-dependent chemolithoautotroph. Our findings rectified the existing knowledge and filled some gaps in episodes of the early evolution of eukaryotes. The further expansion of Heimdallarchaea diversity and their isolation will help to refine evolutionary reconstructions⁴⁸ of eukaryogenesis.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-025-08955-7>.

- Imachi, H. et al. Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).
- Rodrigues-Oliveira, T. et al. Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* **613**, 332–339 (2023).
- Betts, H. C. et al. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).
- Eme, L. et al. Inference and reconstruction of the heimdallarchaeal ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).
- Liu, Y. et al. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).
- Lopez-Garcia, P. & Moreira, D. Open questions on the origin of eukaryotes. *Trends Ecol. Evol.* **30**, 697–708 (2015).
- Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
- Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
- Lu, Z., Zhang, S., Liu, Y., Xia, R. & Li, M. Origin of eukaryotic-like Vps23 shapes an ancient functional interplay between ESCRT and ubiquitin system in Asgard archaea. *Cell Rep.* **43**, 113781 (2024).
- Akil, C. & Robinson, R. C. Genomes of Asgard archaea encode profilins that regulate actin. *Nature* **562**, 439–443 (2018).
- Izore, T., Kureisaite-Ciziene, D., McLaughlin, S. H. & Lowe, J. Crenactin forms actin-like double helical filaments regulated by arcadin-2. *eLife* **5**, e21600 (2016).
- Spang, A., Caceres, E. F. & Ettema, T. J. Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. *Science* **357**, eaaf3883 (2017).
- Williams, T. A., Cox, C. J., Foster, P. G., Szollosi, G. J. & Embley, T. M. Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* **4**, 138–147 (2020).
- Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
- Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).
- Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
- Xie, R. et al. Expanding Asgard members in the domain of Archaea sheds new light on the origin of eukaryotes. *Sci. China Life Sci.* **65**, 818–829 (2022).
- Liu, Y. & Li, M. The unstable evolutionary position of Korarchaeota and its relationship with other TACK and Asgard archaea. *mLife* **1**, 218–222 (2022).
- Da Cunha, V., Gaia, M., Gabelle, D., Nasir, A. & Forterre, P. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13**, e1006810 (2017).
- Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
- Petitjean, C., Deschamps, P., Lopez-Garcia, P., Moreira, D. & Brochier-Armanet, C. Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. *Mol. Biol. Evol.* **32**, 1242–1254 (2015).
- Darling, A. E. et al. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243 (2014).
- Dombrowski, N. et al. Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).
- Fan, L. et al. Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nat. Ecol. Evol.* **4**, 1213–1219 (2020).
- Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
- Viklund, J., Ettema, T. J. & Andersson, S. G. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol. Biol. Evol.* **29**, 599–615 (2012).

27. Ramulu, H. G. et al. Ribosomal proteins: toward a next generation standard for prokaryotic systematics? *Mol. Phylogenet. Evol.* **75**, 103–117 (2014).
28. von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 217 (2019).
29. Hauptfeld, E. et al. Integrating taxonomic signals from MAGs and contigs improves read annotation and taxonomic profiling of metagenomes. *Nat. Commun.* **15**, 3373 (2024).
30. Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
31. Mirdita, M., Steinegger, M., Breitwieser, F., Soding, J. & Levy Karin, E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* **37**, 3029–3031 (2021).
32. Dombrowski, N., Seitz, K. W., Teske, A. P. & Baker, B. J. Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. *Microbiome* **5**, 106 (2017).
33. Zhou, Z., St John, E., Anantharaman, K. & Reysenbach, A. L. Global patterns of diversity and metabolism of microbial communities in deep-sea hydrothermal vent deposits. *Microbiome* **10**, 241 (2022).
34. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).
35. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556 (1997).
36. Kump, L. R. The rise of atmospheric oxygen. *Nature* **451**, 277–278 (2008).
37. Mahendrarajah, T. A. et al. ATP synthase evolution on a cross-braced dated tree of life. *Nat. Commun.* **14**, 7456 (2023).
38. Szollosi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).
39. Sousa, F. L., Neukirchen, S., Allen, J. F., Lane, N. & Martin, W. F. Lokiarchaeon is hydrogen dependent. *Nat. Microbiol.* **1**, 1–3 (2016).
40. Adam, P. S., Borrel, G. & Gribaldo, S. An archaeal origin of the Wood-Ljungdahl H(4)MPT branch and the emergence of bacterial methylophily. *Nat. Microbiol.* **4**, 2155–2163 (2019).
41. Catling, D. C. & Zahnle, K. J. The Archean atmosphere. *Sci. Adv.* **6**, eaax1420 (2020).
42. Tamblyn, R. & Hermann, J. Geological evidence for high H₂ production from komatiites in the Archaean. *Nat. Geosci.* **16**, 1194–1199 (2023).
43. Seitz, K. W. et al. Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat. Commun.* **10**, 1822 (2019).
44. Laso-Pérez, R. et al. Thermophilic archaea activate butane via alkyl-coenzyme M formation. *Nature* **539**, 396–401 (2016).
45. Martin, W. F., Neukirchen, S., Zimorski, V., Gould, S. B. & Sousa, F. L. Energy for two: New archaeal lineages and the origin of mitochondria. *Bioessays* **38**, 850–856 (2016).
46. Mills, D. B. et al. Eukaryogenesis and oxygen in Earth history. *Nat. Ecol. Evol.* **6**, 520–532 (2022).
47. Martin, W. & Müller, M. J. N. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).
48. Giger, G. H. et al. Inducing novel endosymbioses by implanting bacteria in fungi. *Nature* **635**, 415–422 (2024).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

Methods

Sample acquisition and DNA sequencing

We collected 40 sediment samples from six different locations in coastal wetlands across China: Techeng Island (mangrove swamp); Qingmei Port (mangrove swamp); Tongming Port (mangrove swamp); Dongzai Port (mangrove swamp); Changjiang Estuary (salt marsh) and Luchao Port (salt marsh) (Supplementary Table 1 and Supplementary Fig. 1). A peat sampler was employed to collect sediment cores. After collection, sediment samples were sealed in sterile plastic bags, placed immediately in a pre-chilled icebox and transported to the laboratory as quickly as possible, where they were stored at -80°C until DNA extraction. Detailed information of the sampling is presented in the following sections.

Samples from Techeng Island. Two 1-m-deep sediment cores were collected from a mangrove swamp on Techeng Island (Zhanjiang) on November 25, 2018. Three sections with depths of 15–25, 40–45 and 95–100 cm each core were taken in an anoxic glove box. Genomic DNA was extracted from 5–10 g sediment using a PowerSoil DNA Isolation Kit (MoBio Laboratories) according to the manufacturer's instructions. Sequencing libraries were prepared using the NEBNext Ultra DNA Library Prep Kit (New England Biolabs) and sequenced on Illumina HiSeq 2500 instruments at Guangdong MagiGene Technology Corporation, generating approximately 60 Gbp of raw sequencing data (2×150 bp) for each sample.

Samples from Qingmei Port. Two 0.5-m-deep sediment cores were collected from the Qingmeigang Mangrove Reserve (Sanya) in September 2020. Four fractions taken from depths of 0–5, 10–15, 25–30 and 45–50 cm, respectively, for each core were used for DNA extraction. Metagenomic sequencing was carried out on Illumina HiSeq 2500 platform at Suzhou Genewiz Biotechnology Company. Each sample yielded 80–120 Gbp of sequencing data in the form of 2×150 -bp paired-end reads.

Samples from Tongming Port. Three 1-m-deep sediment cores were acquired from a mangrove swamp near Tongming Harbour (Zhanjiang) in June 2021. Samples for DNA extraction were taken from the layers of 15–20 and 45–50 cm of each core. Metagenomic sequencing was performed on Illumina HiSeq 2500 platform at Novogene Biotechnology Company (Tianjin, China), yielding 80–120 Gbp of sequencing data for each sample.

Samples from Dongzai Port. Three sediment cores were obtained from the Dongzhai Port Mangrove Nature Reserve (Haikou) in September 2021. Six samples of each of these cores were taken from depths of 15–20 and 45–50 cm. One additional sample was collected from the 95–100 cm layer of one of the cores. Metagenomic sequencing were conducted as described for samples of Tongming Harbour.

Samples from Changjiang Estuary. Two sediment cores were obtained in November 2018 from the DongTan salt marsh wetland (Shanghai), covered with reeds, located at the estuary of the Yangtze River. Six samples of the two cores were taken from depths of 15–20, 40–55 and 95–100 cm for DNA extraction. Metagenomic sequencing was generated using Illumina HiSeq 2500 by Novogene, yielding 80–120 Gbp of sequencing data for each sample.

Samples from Luchao Port. In November 2021, two sediment cores were obtained from a salt marsh wetland covered with *Spartina alterniflora* near the Luchao Port of Shanghai. Samples for DNA extraction were collected from the layers of 20–25, 40–45 and 95–100 cm of each core. Nucleic acids were sequenced using Illumina HiSeq 2500 by Novogene. Each sample produced 100 Gbp of sequencing data.

Metagenomic assembly and genome reconstruction

A total of 3.54 Tbp of raw reads was obtained from metagenomes generated from the 40 sediment samples. These reads were trimmed using Trimmomatic⁴⁹ (v.0.39) to remove the adaptors and low-quality regions. The resulting clean reads were single-sample de novo assembled using SPAdes⁵⁰ (v.3.15.3) with the parameters: '-k 21,33,55,77,99 -meta'. The assembled scaffolds were used to recruit reads from their own metagenomes and other metagenomes from the same geographical location using Bowtie2 (ref. 51) (v.2.4.4). The SAM files generated were converted to BAM files using Samtools⁵² (v.1.9). The coverage for each of BAM files was then calculated with the jgi_summarize_bam_contig_depth script in MetaBAT2 (ref. 53) (v.2.15). These coverage files were applied as input files for Metabat2 (ref. 53) binning analysis. For each sample, binning was performed using both single-coverage and multi-coverage methods⁵⁴, combined with eight different combinations of specificity and sensitivity parameters: '--maxP 95 or 60, --minS 95 or 60, and --maxEdges 200 or 500. As for a single assembly, an optimized, non-redundant set of bins was chosen by using DAS-Tool⁵⁵ (v.1.1.6). Scaffolds in each bin with divergent GC content or tetranucleotide signatures or coverage profiles were removed with mmgenome⁵⁶ and RefineM⁵⁷. In the end, 11,878 genome bins were obtained from these samples (Supplementary Fig. 1b) and their quality including the completeness, contamination and strain heterogeneity were assessed using CheckM⁵⁸ (v.1.1.3).

Asgard archaeal genome set

A total of 223 high-quality Asgard MAGs (completeness at least 70%, contamination less than or equal to 10%) were reconstructed in this study. An additional 395 high-quality Asgard MAGs plus 15 Njordarchaeales MAGs were downloaded from publicly available databases (NCBI and GTDB¹⁵) as of September 23, 2023. Redundant genomes at species level (ANI at least 95%) were then removed using dRep¹⁴ (v.3.4.3) with parameters '-comp 70 -con 10 -S_an 0.95', resulting in 411 Asgard genome representatives. Of these, 136 were derived from this study, constituting approximately 32% of the total (Supplementary Table 2). Protein coding genes, ribosomal RNA genes and transfer RNA genes in these genomes were identified using Prodigal⁵⁹ (v.2.6.3), Barrnap (v.0.9) (<https://github.com/tseemann/barrnap>) and tRNAscan-SE⁶⁰ (v.2.0.9), respectively.

Phylogenomic analysis

Selection of other archaeal genomes outside Asgardarchaeota. Taxa in the GTDB⁶¹ archaeal reference tree (gtdb_r207_ar53_decorated_fullids_unrooted.tree) were clustered using TreeCluster⁶² with a distance threshold of length1 (parameter: '--threshold 1). After excluding these singleton clusters, the highest quality genome in a cluster was retained. To balance the taxon sampling, 10–15 genomes were selected randomly from each order. For orders with fewer than ten genomes, all available genomes were used. Finally, we obtained 51 DPANN archaea, 47 Euryarchaea and 70 TACK archaea representatives as the outgroup of the following Asgard phylogenies (Supplementary Table 3).

Phylogenomic analyses of previously used archaeal marker sets.

A phylogeny of 53 archaeal marker proteins from the GTDB Toolkit²⁰ (GTDB-Tk) was inferred to examine phylogenetic diversity of Asgard archaea. In addition, to resolve the phylogenetic position of Njordarchaeales in archaea, an additional 12 sets of pre-existing markers that are conserved across archaea were used for inference of trees. Information on these markers is as follows: (1) the RP55 set comprising 55 ribosomal protein sequences^{7,8}, (2) the Dacunha_35uni set consisting of 21 ribosomal proteins and 14 functional proteins¹⁹, (3) the Zare_48uni set consisting of 31 ribosomal proteins and 17 functional proteins⁷, (4) the Spang_udin56ar set comprising 25 ribosomal proteins

and 31 functional proteins²³, (5) the Spang_udin28ar set consisting of ten ribosomal proteins and 18 functional proteins²³, (6) the TomNEE_21 set consisting of nine ribosomal proteins and 12 functional proteins¹³, (7) the Liu_209 set consisting of 45 ribosomal proteins and 164 functional proteins⁵, (8) the RP15 set consisting of 15 ribosomal proteins²⁵, (9) the PV4 set consisting of 30 ribosomal proteins and seven functional proteins²², (10) the Liu_29 set consisting of 23 ribosomal proteins and six functional proteins⁵, (11) the NM200 set consisting of 200 non-ribosomal proteins²¹ and (12) the NM57 set consisting of 57 non-ribosomal proteins⁴ (Supplementary Table 4). These markers were identified in genomes of 411 Asgard archaea and the outgroup by searching against a self-built database composed of all arCOG⁶³ sequences as well as partial sequences from COG⁶⁴, AsCOG⁵, Pfam⁶⁵ and TIGRFAMs⁶⁵, using either BLASTP⁶⁶ or HMMER⁶⁷. For the RP15 set, these ribosomal proteins were adopted if their genes were located on a scaffold harbouring at least five out of 15 ribosomal protein genes. The marker sequences were aligned with MAFFT⁶⁸ (v.7.487, -linsi) and pruned using BMGE⁶⁹ (v.1.12) (-m BLOSUM30). IQ-TREE⁷⁰ (v.2.2.2.6) was used to infer phylogenetic trees under the LG + C60 + F + G + PMSF model. Branch support was assessed with 1,000 ultrafast bootstrap approximations⁷¹ (-B 1,000). Phylogenetic trees were visualized using iTol⁷² (<https://itol.embl.de>). Phylogenomic analyses of all marker sets placed Njordarchaeales within TACK superphylum as a sister clade to Korarchaeota, except for the NM57 dataset.

Construction and phylogenomic analyses of new marker sets.

The expanded diversity of Asgard genomes necessitates new sets of markers that are distributed evenly across Asgard lineages or other archaeal phyla. To further resolve phylogenetic position of Njordarchaeales and eukaryotes within archaea, we integrated these previously reported marker sets used for confidently inferring the archaeal phylogeny^{4,5,13,21} and selected 67 markers (S67) that are conserved across all sampled archaeal genomes. The 67 markers were selected because they were identified in at least 60% of representatives of each of the archaeal clades as well as in 60% of eukaryotic taxa. The S67 marker set was comprised of 39 ribosomal proteins and 28 functional proteins involved in diverse cellular activities (Supplementary Table 4). We used the 67 markers to build four supermatrices: S67, ES67, tES67 and AsES67. The S67 supermatrix was constructed by including all sampled archaeal genomes (411 Asgard archaea, 51 DPANN archaea, 47 Euryarchaea, and 70 TACK archaea representatives), whereas the ES67 supermatrix was built by adding 14 representative eukaryotic genomes into the S67 dataset. These eukaryotic taxa were chosen according to previous studies^{4,7}. The tES67 supermatrix was created by removing DPANN, Euryarchaea, Korarchaeota and Njordarchaeales sequences from the ES67 supermatrix, whereas the AsES67 supermatrix was built by including only 411 Asgard archaea and 14 eukaryotic taxa.

To increase the phylogenetic signal of supermatrices when outgroups were excluded, we relaxed our marker selection criterion and generated two additional marker sets: S150 and S97 (Supplementary Table 4). Both of these were extracted from the aforementioned 13 set of marker proteins^{4,5,7,8,13,15,19,21,22}. The S150 set comprises 150 proteins, including 52 ribosomal proteins and 98 functional proteins, which were selected based on presence of these markers in at least 80% of Asgard archaeal genomes. The S97 set consists of 97 proteins, including 41 ribosomal proteins and 56 functional proteins, which were selected based on their presence of at least 60% of representatives of each of the Asgard clades. The two marker sets were used to build two supermatrices: S150 and S97, by including 411 Asgard archaea and 14 eukaryotic taxa.

Before inferring phylogenies of these marker sets, preliminary trees of each individual marker, inferred using FastTree2 (ref. 73) under the LG + CAT model, were checked manually to identify the correct orthologues for each taxon and to detect any instances of

paralogues. Any identified paralogous sequences were removed manually from the remaining trees. Single-gene trees with paralogous sequences removed have been uploaded to Figshare (<https://figshare.com/s/6e523322b0b647b91dda>)⁷⁴. In addition, by excluding markers that failed to recover archaea and bacteria as reciprocally monophyletic domains, we obtained 47, 60 and 99 markers (S47, S60 and S99) from the S67, S97 and S150 marker sets, respectively (Supplementary Table 4). Detailed analyses of these marker sets are provided in the Supplementary Information. ML trees of these amino acid supermatrices were inferred using IQ-TREE⁷⁵ with the LG + C60 + F + G + PMSF model. The model was selected because it can capture variations in the substitution or replacement process across sites. The robustness of the resulting trees was assessed with 1,000 ultrafast bootstrap approximations⁷¹ (-B 1,000). Bayesian inferences were performed using PhyloBayes-MPI⁷⁶ (v.1.9) with CAT + GTR after the alignments were recoded into four categories using the SR4 scheme³⁴. Two independent Markov chains were run until a sufficient effective sample size (greater than 300) was achieved. For consensus tree reconstruction, the first 5,000 cycles (for the run with 20,000 iteration times) or the first 10,000 cycles (for the run with 50,000 iteration times) were discarded as burn-in. In all Bayesian inferences, the two chains did not converge, probably due to the large size of the supermatrices and computational limitations.

To examine effect of mutationally saturated or rate-biased sites on tree reconstruction^{24,26,77}, we tracked the bootstrap support values of branches of interest. The evolutionary rate of sites (evolving-rate score) was estimated using IQ-TREE⁷⁰ with the empirical Bayesian algorithm⁷⁸ (-rate). The sites were classified into ten categories from the fastest to the slowest evolving. A series of alignments were generated by removing 10% to 90% of the data in a stepwise fashion. The resulting alignments were used to compute 1,000 ultrafast bootstrap⁷¹ support values in IQ-TREE⁷⁵, using the LG + C60 + F + G + PMSF model. In the end, the bootstrap support for each bipartition of interest was calculated and tracked as the fastest-evolving sites were progressively removed.

Assessment of Njordarchaeales representative MAGs. The taxonomic profiles of contigs of Njordarchaeales MAGs were determined using two state-of-the-art tools: CAT^{28,29} and MMseqs2 (refs. 30,31), by comparison with TACK and Asgard archaeal sequences from the GTDB database⁷⁹ (r220) excluding Njordarchaeales-related sequences. CAT is a tool for taxonomic annotation of contigs based on protein homologies to a reference database, while MMseqs2 determines the taxonomic identity of contigs through a weighted voting mechanism. The two tools both rely on a well-defined reference database. Given that the position of Njordarchaeales moved between TACK superphylum and Asgardarchaeota in species trees, we retrieved representative genomes of TACK and Asgard archaea only from the GTDB¹⁵ database (r220) to construct the reference database. This strategy aims to minimize the impact of other archaeal sequences on taxonomic classification. As shown in Fig. 2a,b, 20–73% of the contigs in each MAG were not assigned to a specific classification, probably indicating that these contigs contain sequences that are too divergent from those in the database.

Metagenomic samples from which Njordarchaeales MAGs were reconstructed were downloaded from publicly available databases^{4,32,33}. These metagenomes were generated from deep-sea hydrothermal vent sediments. Metagenomic reads were mapped to these Njordarchaeales MAGs using Bowtie2 (ref. 80) (v.2.3.5). The resulting SAM files were converted to BAM files with Samtools³² (v.1.9), which were used to create an input index for Anvi'o⁸⁰. Coverage and GC content of contigs were recorded. Contigs of these MAGs were clustered according to their sequence composition and coverage of contigs. Contigs originating from the same organism typically exhibit similar sequence coverage. Therefore, those with divergent coverage profiles are often considered

Article

as potential contaminants^{81,82}. These contaminating contigs are particularly prone to displaying differential coverage patterns across different metagenomic samples⁸¹. In this study, four Njordarchaeales MAGs were completely matched by reads from several metagenomes generated from similar habitats. The average coverages of contigs in the clusters differed across metagenomes.

Phylogenetic divergence time estimation. Molecular dating analysis was carried out with the MCMCtree program in PAML⁸³ using the WAG model. The analysis was run iteratively until convergence was achieved. Seven nodes in the phylogenetic tree were calibrated using four types of constraint⁸⁴: (1) the Archaeal root constraint was set at 4.29–3.80 Ga, (2) the chitin age had a minimum constraint of 1.58 Ga, (3) the oxygen age had a maximum constraint of 2.32 Ga and (4) the HGT from Viridiplantae to Thaumarchaeota was constrained to 1.49–0.75 Ga. Further details are provided in Supplementary Table 5.

Ancestral metabolic reconstruction. For analyses of amalgamated likelihood estimation^{38,85}, a set of archaeal genomes with more than 80% completeness and less than 5% contamination (mean completeness of 88.5% and mean contamination of 2.7%) was selected, comprising 235 Asgard archaea, 47 Euryarchaea and 60 TACK archaea representatives (Supplementary Table 2 and 3). These genomes had a mean quality score of 75%, with the lowest quality value⁵⁷ being 51%. All predicted protein sequences from these genomes were annotated by searching against arCOGs⁶³ and nr database using BLASTP⁸⁶ (E-value less than 1×10^{-5}) or Pfam⁶⁵ and TIGRFAMs⁸⁷ databases using HMMER⁸⁸ with a GA bitscore cutoff. The SwissProt database⁸⁹ was further used to validate protein functions. Protein sequences with identical functional descriptions were clustered to infer individual gene trees. Sequences were aligned and pruned using MAFFT⁶⁸ (-linsi) and BMGE⁶⁹ (-m BLOSUM30). Gene tree profiles were then inferred using FastTree2 (ref. 73) with the LG + CAT model. Subsequently, the species tree of the S47 dataset, along with the gene tree profiles, were applied to perform gene–tree species tree reconciliations using ALEobserve (v1.0) and ALEml_undated. Events of loss, transfer, origination or duplication as well as presence/absence of gene (copies) were considered only if they had a raw reconciliation frequency of at least 0.1. The copy number (proteome size) and the events are summarized in Supplementary Table 6 and 8, and visualized on the branches of the species tree (Supplementary Fig. 18). The resulting data were then used for metabolic reconstruction. To support the inferences of ALE, gene frequencies for key metabolic pathways across Asgard archaea are provided in Supplementary Table 7.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All sequence data are archived in the NCBI database under BioProject ID PRJNA1162170. All raw data underlying phylogenomic analyses (raw and processed alignments and corresponding phylogenetic trees) are available at Figshare (<https://figshare.com/s/6e523322b0b647b91dda>)⁷⁴.

49. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
50. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
51. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357 (2012).
52. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
53. Kang, D. W. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

54. Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat. Methods* **20**, 1170–1173 (2023).
55. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
56. Albertsen, M. et al. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
57. Parks, D. H. et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
58. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
59. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinf.* **11**, 119 (2010).
60. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
61. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
62. Balaban, M., Moshiri, N., Mai, U., Jia, X. & Mirarab, S. TreeCluster: clustering biological sequences using phylogenetic trees. *PLoS ONE* **14**, e0221068 (2019).
63. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. *Life (Basel)* **5**, 818–840 (2015).
64. Galperin, M. Y. et al. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).
65. Finn, R. D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–D230 (2014).
66. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
67. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled–coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
68. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
69. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
70. Minh, B. Q., Trifunopoulos, J., Schrepf, D., Schmidt, H. & Lanfear, R. IQ-TREE version 2.0: tutorials and manual phylogenomic software by maximum likelihood. *Nucleic Acids Res.* **44**, W232–W235 (2022).
71. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
72. Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).
73. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2: approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
74. Zhang, J. et al. Phylogenomic analyses with expanded taxon sampling reveal deep origin of eukaryotes outside Heimdallarchaeia. *Figshare* <https://figshare.com/s/6e523322b0b647b91dda> (2025).
75. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
76. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615 (2013).
77. Munoz-Gomez, S. A. et al. Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat. Ecol. Evol.* **6**, 253–262 (2022).
78. Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781–1791 (2004).
79. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
80. Eren, A. M. et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
81. Chen, L. X., Anantharaman, K., Shaiber, A., Eren, A. M. & Banfield, J. F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).
82. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
83. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
84. Yang, Y. et al. The evolution pathway of ammonia-oxidizing archaea shaped by major geological events. *Mol. Biol. Evol.* **38**, 3637–3648 (2021).
85. Szollosi, G. J., Davin, A. A., Tannier, E., Daubin, V. & Boussau, B. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140335 (2015).
86. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
87. Haft, D. H., Selengut, J. D. & White, O. The TIGRFAMs database of protein families. *Nucleic Acids Res.* **31**, 371–373 (2003).
88. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
89. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).

Acknowledgements H.-P.D. and L.-J.H. were supported by the Chinese National Key Programmes for Fundamental Research and Development (2024YFF0808800), the National Natural Science Foundation of China (42476110, 42030411, 41971125) and the Ocean Negative Carbon Emissions (ONCE) Program. M. Li was supported by the National Natural Science Foundation of China (32393971, 32225003, 32393970, 92251306), Guangdong Major Project of Basic and Applied Basic Research (2023B0303000017) and Shenzhen University 2035 Program for Excellent Research (2022B002). X.F. was supported by Guangdong Basic and Applied Basic Research (2025A1515012817).

Author contributions This study was conceived by H.-P.D. with input from J.Z. and L.-J.H. Metagenomic assembly, binning and phylogenetic analyses were conducted primarily by J.Z. X.F. provided bioinformatics support, with input from M. Li and Y.L. H.-P.D., L.-J.H. and M. Liu

contributed data. H.-P.D. wrote the manuscript and Supplementary Information, with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

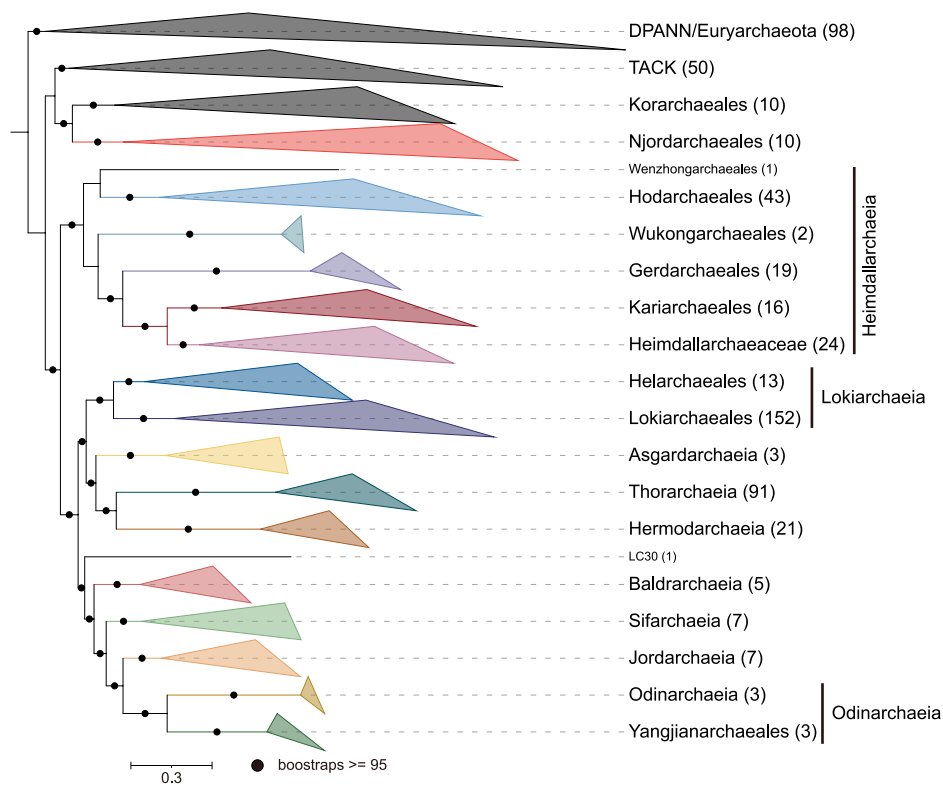
Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41586-025-08955-7>.

Correspondence and requests for materials should be addressed to Li-Jun Hou or Hong-Po Dong.

Peer review information *Nature* thanks William Martin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

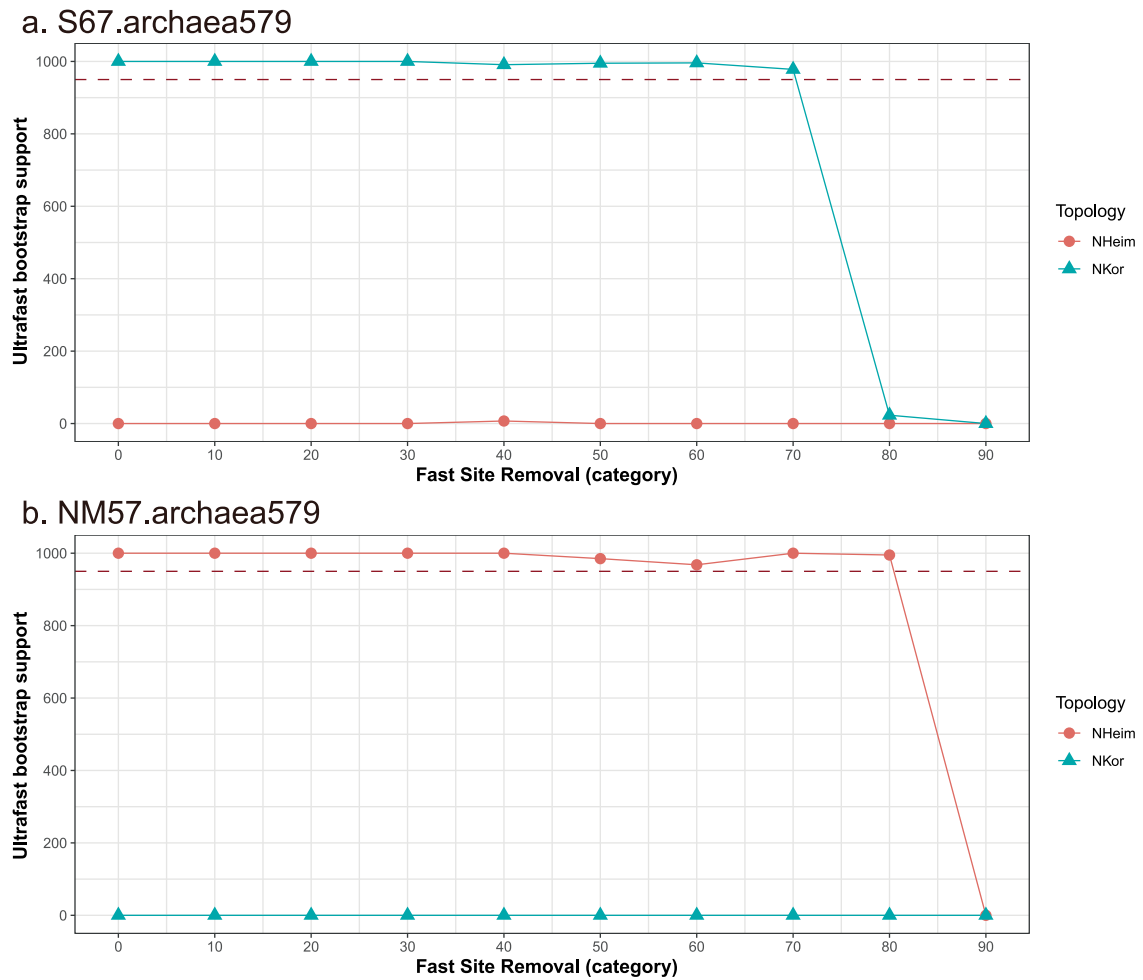
Reprints and permissions information is available at <http://www.nature.com/reprints>.

S67.archaea579



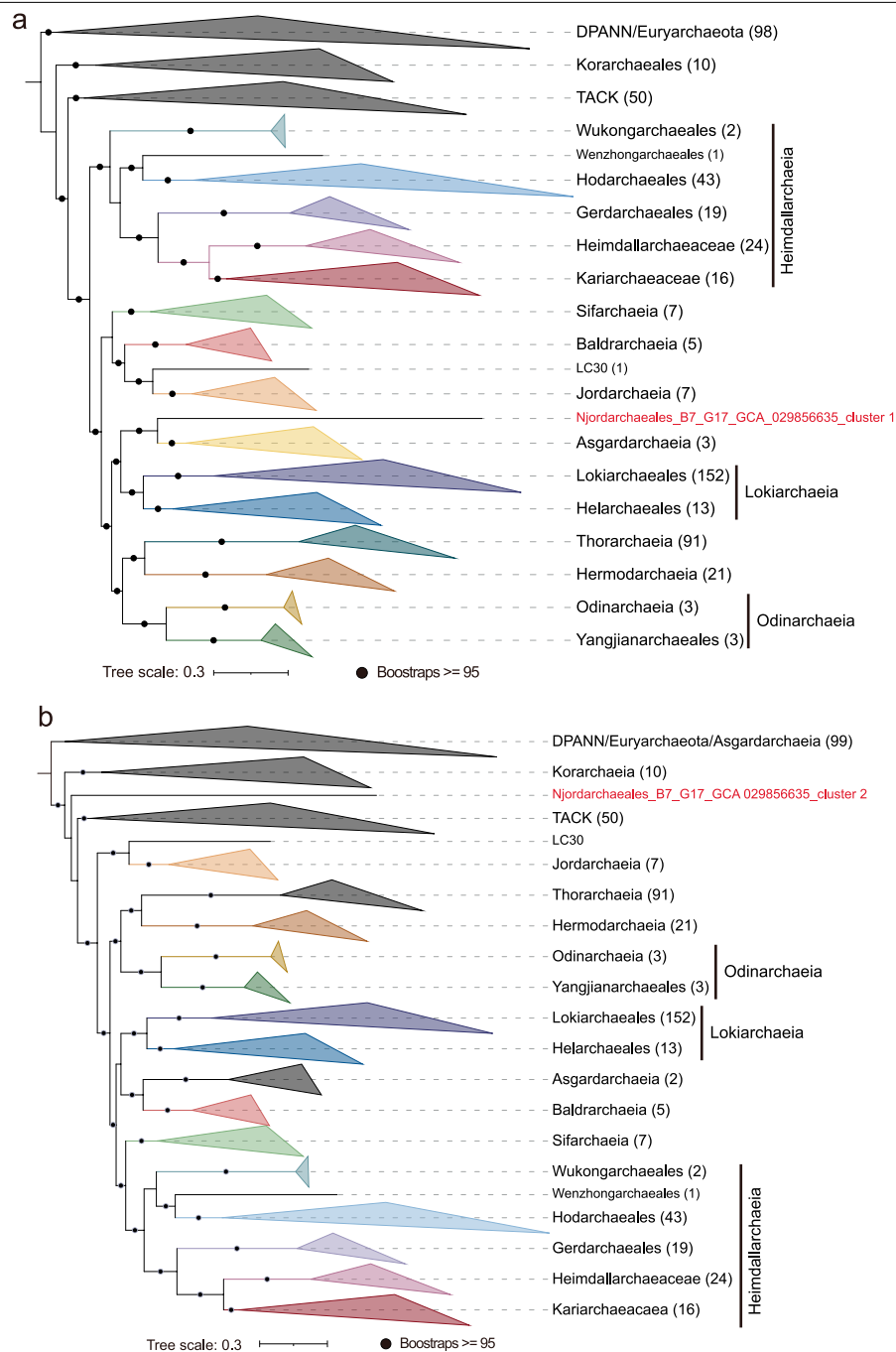
Extended Data Fig. 1 | The maximum likelihood phylogenomic analysis of the S67 dataset. The tree was inferred using IQ-TREE under the LG + C60 + F + G + PMSF model, based on 579 archaeal taxa (11,860 sites, 67 concatenated

proteins, 579 taxa). This tree was rooted to DPANN and Euryarchaeota. Bootstrap support values $\geq 95\%$ are represented by black dots. The scale bar denotes the average expected number of substitutions per site.



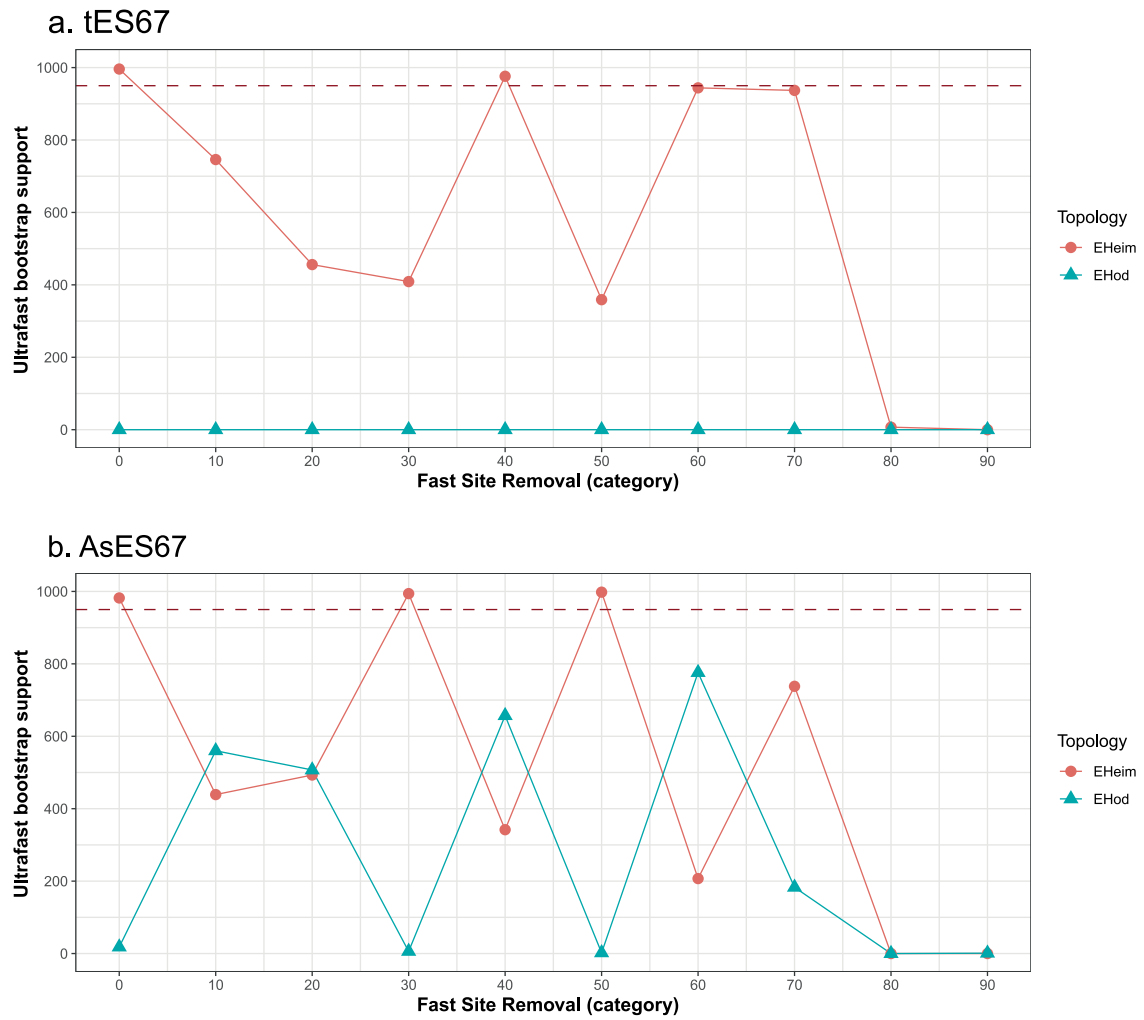
Extended Data Fig. 2 | Evolution of ultrafast bootstrap support for either the monophyly of Njordarchaeales and Heimdallarchaeia (NHeim, red line) or the monophyly of Njordarchaeales and Korarchaeia (NKor, blue line). The bootstrap values were obtained from phylogenomic trees inferred from

the S67 (a) and NM57 (b) datasets, based on 579 archaeal taxa, with the fastest-evolving sites progressively removed. The trees were inferred using IQ-TREE under the LG + C60 + F + G + PMSF model.



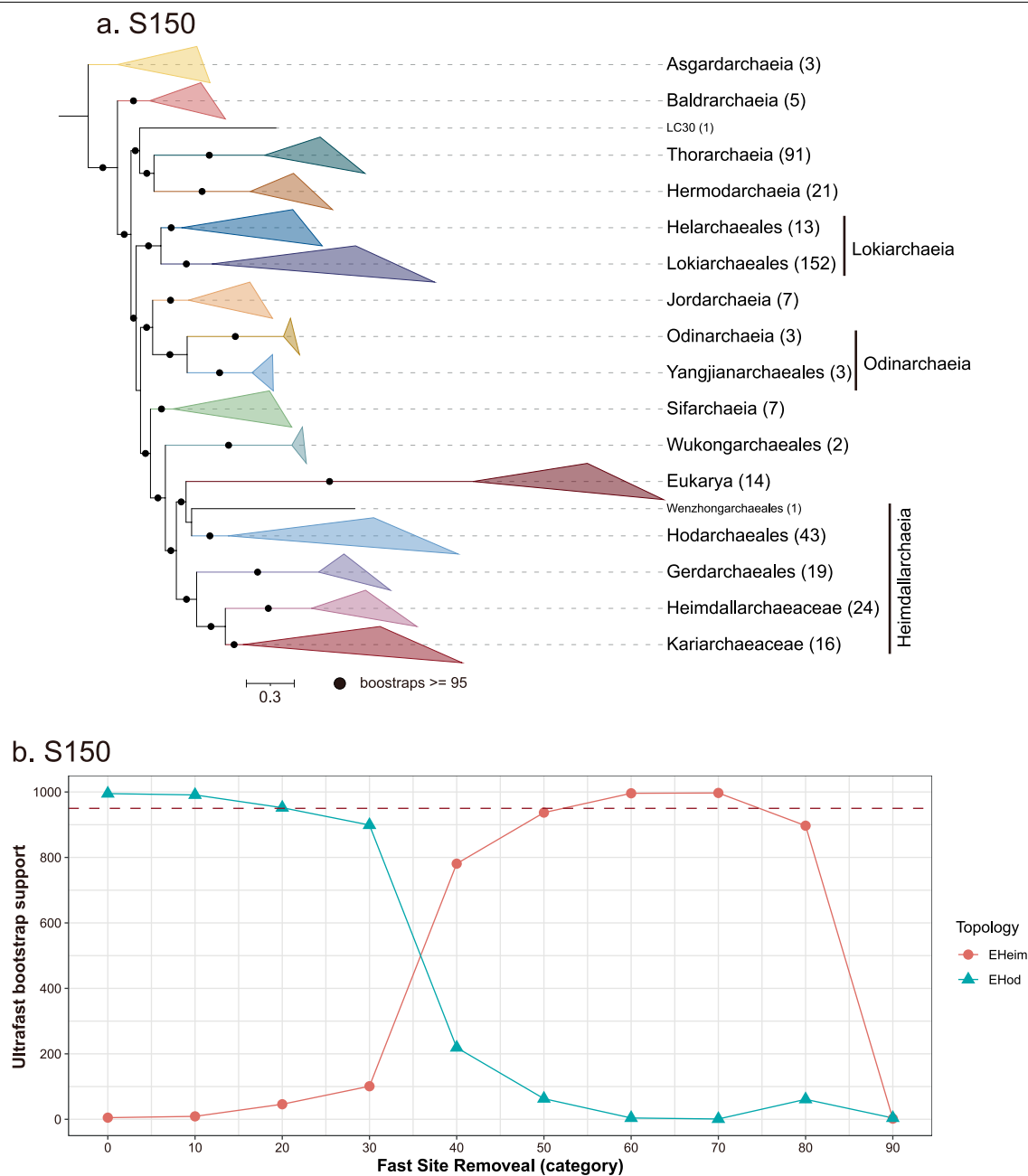
Extended Data Fig. 3 | The maximum likelihood phylogenomic analysis of NM57 marker set, showing the position of contigs in two larger clusters of Njordarchaeales B7_G17_GCA_029856635. Contigs of Njordarchaeales B7_G17_GCA_029856635 were grouped into three clusters (Fig. 2c). **a**, Phylogenetic position of contigs in cluster 1. **b**, Phylogenetic position of contigs in cluster 2.

The trees were inferred using IQ-TREE under the LG + C60 + F + G + PMSF model and rooted to DPANN and Euryarchaeota. Bootstrap support values $\geq 95\%$ are represented by black dots. The scale bar denotes the average expected number of substitutions per site.



Extended Data Fig. 4 | Evolution of ultrafast bootstrap support for the monophyly of either eukaryotes and Heimdallarchaeia (EHeim, red line) or the monophyly of eukaryotes and Hodarchaeales (EHod, blue line). The bootstrap values were obtained from the phylogenies inferred from the tES67

dataset (a) and AsES67 dataset (b), as the fastest-evolving sites were progressively removed. The trees were inferred using IQ-TREE under the LG + C60 + F + G + PMSF model.



Extended Data Fig. 5 | Phylogenetic analyses and site-exclusion treatment of the S150 dataset for 411 Asgard archaeal taxa and 14 eukaryotic taxa. **a.** The maximum likelihood phylogenomic analysis based on the S150 dataset (32,277 sites, 150 concatenated proteins, 425 taxa). The trees were inferred using IQ-TREE under the LG + C60 + F + G + PMSF model. Bootstrap support values $\geq 95\%$ are represented by black dots. The scale bar denotes the average

expected number of substitutions per site. **b.** Evolution of ultrafast bootstrap support for the monophyly of either eukaryotes and Heimdallarchaeia (EHeim, red line) or the monophyly of eukaryotes and Hodarchaeales (EHod, blue line), in the phylogenies inferred from the S150 dataset. The trees were inferred using IQ-TREE under the LG + C60 + F + G + PMSF model.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Bowtie2 v2.4.4; MetaBAT2 v2.15; DAS Tool v1.1.6; Trimmomatic v0.38; metaSPAdes v3.15.3; SAMtools v1.9
Data analysis	Published software used for data analysis include CheckM v1.1.13; Prodigal v2.6.3; tRNAscan-SE v2.0; psiblast v2.12.0; MMseqs2; MUSCLE v5.0.1; FastTree v2.1.11; Barrnap v0.9; mafft v7.487; BMGE v1.12; IQ-Tree v2.2.2.6; Anvi'o v8.0; PhyloBayes-MPI v1.9; GTDBtk v2.1.0; HMMER v3.3.2; Bowtie2 v2.4.4; SAMtools v1.9; MEGA X; PAML v4.10.6; ALE v1.0; Interproscan 5.52-86.0; kofam_scan v1.3.0; dRep V3.4.3; CAT V6.0.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Asgard archaeal genomes generated in this study are deposited in NCBI under BioProject ID PRJNA1162170. All raw data underlying phylogenomic analyses (raw and processed alignments and corresponding phylogenetic trees) have been deposited on Figshare (<https://figshare.com/s/6e523322b0b647b91dda>).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

Use the terms *sex* (biological attribute) and *gender* (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.

Reporting on race, ethnicity, or other socially relevant groupings

Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

Ethics oversight

Identify the organization(s) that approved the study protocol.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

The number of lineages used in phylogenomic analyses is based on a balance between diversity and representativeness of species and computational resources.

Data exclusions

No data was excluded

Replication

Robustness and reliability of all maximum likelihood trees were evaluated using 1000 ultrafast bootstrap replicates.

Randomization

Randomness is not applicable in phylogenetic and comparative genomic analyses.

Blinding

Blinding is not applicable in phylogenetic and comparative genomic analyses.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.